# Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings

**Philippa Shoemark**[*♠♢], **Farhana Ferdousi Liza**[*♠♡], **Dong Nguyen**[♠♢],
**Scott A. Hale**[♠†], **Barbara McGillivray**[♠‡]
[♠]The Alan Turing Institute, UK, [♢]School of Informatics, University of Edinburgh, UK,
[♡]School of Computing, University of Kent, UK, [†]Oxford Internet Institute, University of Oxford, UK,
[‡]Faculty of Modern and Medieval Languages, University of Cambridge, UK
p.j.shoemark@ed.ac.uk, fl207@kent.ac.uk, dnguyen@turing.ac.uk,
scott.hale@oii.ox.ac.uk, bmcgillivray@turing.ac.uk

## Abstract

Word embeddings are increasingly used for the automatic detection of semantic change; yet, a robust evaluation and systematic comparison of the choices involved has been lacking. We propose a new evaluation framework for semantic change detection and find that (i) using the whole time series is preferable over only comparing between the first and last time points; (ii) independently trained and aligned embeddings perform better than continuously trained embeddings for long time periods; and (iii) that the reference point for comparison matters. We also present an analysis of the changes detected on a large Twitter dataset spanning 5.5 years.

## 1 Introduction

Semantic change, i.e., the change in the meanings of words, is inherent in language. A new meaning for a word can be added to the original one, become more or less prevalent, or even replace a former meaning (see Koch, 2016). An example is *lit*, which has gained a new sense of 'exciting' or 'awesome', via the extension of its long-established use as slang for 'intoxicated' to describe the vibrant environment in which acts of becoming intoxicated often occur.[1]

Automatically measuring semantic change can discover changes that would not be apparent from manual inspection. It can also facilitate the investigation of mechanisms driving semantic changes, e.g., how these changes are affected by language-internal and social factors. Moreover, there are direct benefits to applications, such as the detection of meaning shifts in polarized words to update sentiment lexicons and the detection of emerging word meanings to update dictionaries.

Word embeddings are increasingly used for automatic semantic change detection (Kutuzov et al., 2018). Words are mapped to low-dimensional vectors, and the semantic change of a word is then measured by comparing its vectors across time periods. Although word embeddings have emerged as one of the most popular approaches to measuring semantic change, researchers are faced with various decisions, including whether to train embeddings independently or continuously, which metric to use to measure change between two time periods, and which ranking approach to use for comparing semantic change candidates.

A major challenge in developing semantic change detection systems is obtaining ground truth data (Kutuzov et al., 2018), which has so far prevented a systematic evaluation of different approaches. Many studies rely on hand-picked examples (e.g., Wijaya and Yeniterzi, 2011; Rodda et al., 2017) or human judgements (e.g., Tredici et al., 2018). Some studies have performed evaluations based on dictionary data (e.g., Cook et al., 2014; Basile and McGillivray, 2018), manual annotation of dictionary senses in corpora (McGillivray et al., 2019), and manual annotation of word types (Kenter et al., 2015), but this approach is not well-suited for recent, yet-to-be-recorded changes.

In this paper, we present a new framework to evaluate semantic change detection systems (Section 5.1). We model multiple semantic change scenarios and compare the impact of different choices that are typical when using word embeddings to analyse semantic change (Section 5.2). Our framework is not specific to the use of word embeddings and can also support the evaluation of other approaches not considered in this paper.[2] We

*These authors contributed equally to the study.

[1]"What Does 'Lit' Mean?" Retrieved from https://www.merriam-webster.com/words-at-play/lit-meaning-origin. Accessed 2018-10-05.

[2]The dataset and all code for this paper is available at https://github.com/alan-turing-institute/room2glo.

then apply the approaches to 5.5 years of Twitter data and provide an in-depth analysis of the top-ranked semantic change candidates (Section 6).

## 2 Related Work

There has been increasing interest in automatic semantic change detection (Tang, 2018; Kutuzov et al., 2018), using methods ranging from neural models to Bayesian learning (e.g., Frermann and Lapata, 2016), Temporal Random Indexing (e.g., Basile and McGillivray, 2018) and dynamic topic modelling (e.g., Blei and Lafferty, 2006). Word embeddings have been especially popular (e.g., Dubossarsky et al., 2017; Hamilton et al., 2016b), and recently Bamler and Mandt (2017) and Rudolph and Blei (2018) explored dynamic embeddings for semantic change detection by training a joint model over all time periods.

Most previous work analysed corpora spanning long time periods (e.g., a few centuries), such as the Google Books Ngrams corpus and the Corpus of Historical American English (e.g., Hamilton et al., 2016b). Recently short-term semantic changes have been studied, for example in Amazon Reviews (Kulkarni et al., 2015), scientific papers (Rudolph and Blei, 2018), news articles (Tang et al., 2016; Yao et al., 2018), and the UK Web Archive (Basile and McGillivray, 2018).

In this paper, we focus on social media: in particular, on Twitter. Semantic change in social media has only been lightly explored, with studies on Twitter (Kulkarni et al., 2015), the VKontake social network (Stewart et al., 2017), and Reddit (Tredici et al., 2018). In comparison to these studies, our data covers a longer time period and our evaluation more deeply explores the various choices involved in semantic change detection.

Much of the previous work on semantic change discovery has relied on qualitative evaluations of small samples from the output, case studies of a few well-known historical changes (e.g., Kim et al., 2014; Hamilton et al., 2016a,b; Stewart et al., 2017), or attested changes extracted from dictionaries (e.g., Rohrdantz et al., 2011; Cook et al., 2014; Basile and McGillivray, 2018). Some evaluations have been based on related tasks for which performance is expected to correlate, such as classifying the time period a text snippet belongs to (Mihalcea and Nastase, 2012) or predicting real-world events (Kutuzov et al., 2017).

Here we look for meaning changes over a short, recent time period. There is little existing literature on words that have undergone meaning change within the relevant time-frame, and language on social media is not always fully reflected in general language dictionaries. Moreover, even if we were able to obtain a substantial list of attested meaning changes, a system might still discover other valid meaning change candidates. Unfortunately, determining the validity of semantic change candidates is time-consuming, labour-intensive, and subjective; so, building on prior approaches (e.g., Kulkarni et al., 2015; Rosenfeld and Erk, 2018; Nguyen and Eisenstein, 2017), we introduce a new synthetic evaluation framework for semantic change detection.

Synthetic evaluation is especially important for short, recent time periods given the lack of other resources for evaluation, but it is also valuable for longer periods to detect hitherto unknown changes. Moreover, phenomena like seasonal trends are more likely to interfere with semantic change detection for short time periods, making this a challenging use case for semantic change detection. At the same time, it is an important use case in order to advance semantic change detection for contemporary data to be used to update lexicons, sentiment/polarity ratings, and other language resources.

## 3 Data

We collected tweets from Twitter's 'statuses/sample' streaming API endpoint from January 1, 2012, to June 30, 2017. There are a few minor gaps in our data due to occasional data collection issues. Most are a few minutes or at most a day, but one gap spans January and February 2015. Overall, our dataset consists of over 7 billion tweets sent during 1,889 days.

We use the Compact Language Detector version 2 (CLDv2),[3] following guidance from Graham et al. (2014), and we discard any tweets for which CLD detects less than 90% of the text to be in English, resulting in roughly 2.5 billion tweets. The remaining tweets are then lowercased, and usernames, urls, and non-alphanumeric characters (except emoji and hashtags) are removed. The text is then tokenized on whitespace. Digit-only tokens are replaced with '<NUM>'. Finally, we discard tweets that are duplicated within a given month, as tweets which are re-tweeted or copied verba-

[3] https://github.com/CLD2Owners/cld2

tim many times are not independent language samples and may exert undue influence on embeddings (Mikolov et al., 2018). Our final dataset consists of 1,696,142,020 tweets and 20,273,497,107 tokens.

## 4 Methods

Following the approach introduced by Kim et al. (2014) and adopted by Hamilton et al. (2016b) and others, we divide our dataset into discrete time periods, and for each time period $t$ we compute word embeddings, representing each word $w$ by a $d$-dimensional vector. We then compare the embeddings between different time periods to measure the semantic change of words. We use monthly bins, but the approach is applicable to time periods of any length, provided there is sufficient data in each bin to train quality embeddings.

### 4.1 Training Word Embeddings

We train word embeddings using gensim's (Řehůřek and Sojka, 2010) implementation of the continuous bag of words (CBOW; Mikolov et al., 2013) model.[4] Two evaluation tasks (word similarity using the dataset Wordsim353[5] and word analogy using the word test dataset[6]) were used to tune four hyperparameters, resulting in 200 dimensions, a window size of 10, 15 iterations, and a minimum frequency of 500 (per time-step). For all other hyperparameters we use gensim's default values.

### 4.2 Comparable Embeddings

To compare embeddings for a word between two time-points, the embeddings need to be in the same coordinate axes. We experiment with three approaches: (1) Training continuously by initializing the embeddings for a given time-step $t$ with the embeddings trained at the previous time-step $t-1$ (e.g., Kim et al., 2014); (2) Training embeddings for each time-step independently and post-hoc aligning them (e.g., Hamilton et al., 2016b; Kulkarni et al., 2015) using orthogonal Procrustes (as used by Hamilton et al., 2016b); and (3) combining continuous training and post-hoc alignment (as in Stewart et al., 2017).

### 4.3 Measuring Semantic Change

We compare two measures for quantifying a word's semantic change between two time points. The first is the cosine distance, a common approach in previous work (Hamilton et al., 2016b; Stewart et al., 2017; Dubossarsky et al., 2017; Kim et al., 2014). The second measure, introduced by Hamilton et al. (2016a), is based on comparing the neighbourhoods of the embeddings. For each time-step $t$, we first find the ordered set of word $w$'s $k$ nearest neighbours, based on cosine similarity. Following Hamilton et al. (2016a), we set $k = 25$. For any two time-steps, we then take the union $S$ of the two nearest neighbour sets and create a second-order vector $v_t$ where each entry $v_t^{(i)}$ contains the cosine similarity of target word $w$ to neighbouring word $S^{(i)}$ at time $t$. We then measure the cosine distance between these two second-order vectors.

### 4.4 Ranking Semantic Change Candidates

Our goal is not only to measure semantic changes for pre-selected words, but to identify which words out of the entire vocabulary have undergone the greatest or most significant semantic change. We compare several approaches to generating ranked lists of the 'most changed' words. The first only measures the change between two time-steps. The remaining approaches consider the whole time series. For the approaches that use the whole time-series, we limit the semantic change candidates to words that occur at least 500 times in at least 75% of the time-steps, simply condensing a word's time-series if there are gaps.

**Two-step approach** We first measure each word's semantic change between just two pre-selected time-steps (in this study, the first and final time-steps). This simple approach has been used in previous work, such as Kim et al. (2014).

**Change-point detection** Following Kulkarni et al. (2015), we choose one time-step $t_0$ as a reference and compute semantic change scores for each word with respect to $t_0$ at every other time-step $t_i$. Then, for each word $w$ and each time-step $t_i$, we compute a *mean-shift* score by partitioning $w$'s time series of semantic change scores at $t_i$, and

[4] We only report results using CBOW in this paper. We found similar trends when using the skip-gram model, which has been used in previous works on semantic change (e.g., Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016b; Stewart et al., 2017; Tredici et al., 2018).

[5] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

[6] http://www.fit.vutbr.cz/~imikolov/rnnlm/word-test.v1.txt

calculating the difference between the means of the scores in the two partitions. Following Kulkarni et al. (2015), we use Monte Carlo permutation tests to estimate the statistical significance of mean-shift scores, and take the time-step with the lowest estimated p-value as the change point. Words are first sorted in descending order of the mean-shift scores of their estimated change points and then in ascending order of their p-values.

Kulkarni et al. standardized each word's cosine score for a given time-step relative to the mean score across all words at that time-step. This is meant to help control for corpus artefacts, e.g., shifting sampling biases over time, but its impact has not been demonstrated yet. We compare the results of ranking words without standardization (*raw* scores) and with standardization ($z$ scores).

**Global trend detection** We also compare three approaches to detect global trends in the same time series of semantic change scores as the change point detection methods. The first approach is fitting a linear regression model $d_i = \alpha + \beta t_i + \epsilon_i$, where $d_i$ is semantic change distance scores, $t_i$ is time periods $\{1, ..., n\}$, and $\epsilon_i$ is error. We rank the words based on their absolute $\beta$ values (slopes), which gives the semantic change per time period under the assumption of a linear relationship.

We also experiment with two correlation measures: Pearson's ($r$) and Kendall's rank ($\tau$) correlation coefficients. In contrast to linear regression and Pearson's correlation coefficient, Kendall's tau is non-parametric and resistant to outliers (Kendall, 1948). It is therefore often used for measuring trends in time series and change point detection (Quessy et al., 2013). We rank the words based on the absolute values of $\tau$ and $r$.

## 5 Synthetic Evaluation

To systematically compare the different methodological choices we introduce a new synthetic evaluation framework. We create seven schemas for how a word's distributional statistics may change. Three of these model scenarios in which a semantic change occurs, but crucially the remaining four model scenarios that we would *not* wish to classify as semantic change. Our framework builds on previous approaches that have modelled one type of semantic change—either a word gaining an additional sense (Kulkarni et al., 2015) or a word's original sense being completely replaced (Rosenfeld and Erk, 2018). Furthermore, although most work has focused on *recall*, our framework can also test *precision*, i.e., the ability to distinguish the injected changes from noise.

### 5.1 Dataset Construction

We first randomly sample 10% of the tweets from a single month from the middle of our empirical dataset (Dec. 2014). We then draw 66 random 70% samples with replacement from this sample. These 66 samples represent a dataset of 66 months (5.5 years) in which no semantic changes occur, but words' distributional statistics still vary from month to month due to sampling noise. This differs from, e.g., Kulkarni et al. (2015), who used a series of exact duplicates of an initial set of documents. Finally, we inject controlled changes by inserting made-up 'pseudowords', carefully changing their frequencies and co-occurrence distributions throughout the time series.

Our procedure for inserting pseudowords is as follows: we split the real words that occur in our empirical data for December 2014 into five equally sized frequency bins. For each pseudoword $\rho$ that we insert, we choose a frequency bin. To represent one of the senses of $\rho$, we sample a real word $w$ from the relevant frequency bin. For each synthetic month $m$, we insert $\rho$ replacing each token of $w$ with success probability $p_{(\rho,w,m)}$.

For example, we might insert one pseudoword replacing the instances of the word 'pudding' with a fixed probability throughout the whole time series, and then insert this same pseudoword replacing the instances of the word 'neon' with increasing probability over time. This would model a word that initially has a meaning related to 'pudding', but which then acquires a new sense related to 'neon'. We use seven different schemas: three model different kinds of semantic change (C1–C3), and four model ephemeral changes that we aim to avoid (D1–D4): see Figure 1.

C1: **Description:** This schema models a word that gradually acquires a new sense over time while retaining its original sense (e.g., *snowflake*, *lit*). This corresponds to what Koch (2016, 24) calls 'innovative meaning change' and Tahmasebi et al. (2018, 35) calls 'novel word sense'.

**Procedure:** Sample one real word $w_1$ to represent the original pseudosense[7] of the pseudoword $\rho$ and another real word $w_2$ to rep-

[7] A single pseudosense may in practice correspond to mul-

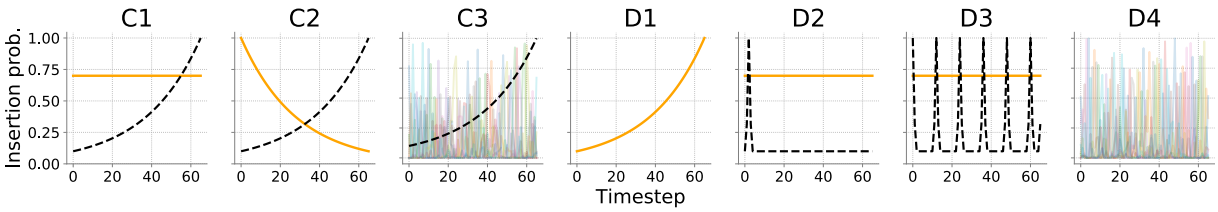


Figure 1: Illustration of our seven schemas for inserting pseudowords into the synthetic dataset. Each line represents a different pseudosense. Lines chart the probability of inserting a pseudoword token replacing a token representing the relevant pseudosense, as a function of 'time'. We vary whether the success probabilities change linearly or logarithmically and the time-steps at which the changes begin and end.

resent its new pseudosense. For each token of $w_1$ that occurs in the synthetic dataset for month $m$, insert a token of $\rho$ replacing it with success probability $p_{(\rho,w_1,m)}$, which remains constant throughout the time series. Insert $\rho$ replacing each token of $w_2$ with success probability $p_{(\rho,w_2,m)}$, which starts low and gradually increases over time.

C2: **Description:** This schema models a word that gradually acquires a new sense over time while its original sense gradually falls out of use (cf., *silly*, which originally meant 'happy' or 'lucky' and now means 'foolish'). This corresponds to the full cycle of genesis and disappearance of lexical polysemy as described by Koch (2016, 25), i.e., an 'innovative meaning change' and a 'reductive meaning change'.

**Procedure:** Sample one real word $w_1$ to represent the original sense of the pseudoword $\rho$, and another real word $w_2$ to represent its new sense. For each token of $w_1$ that occurs in the synthetic dataset for month $m$, insert a token of $\rho$ replacing it with success probability $p_{(\rho,w_1,m)}$, which starts relatively high and gradually decreases over time. Insert $\rho$ replacing each token of $w_2$ with success probability $p_{(\rho,w_2,m)} = 1 - p_{(\rho,w_1,m)}$.

C3: **Description:** This schema models a word with many senses, different random subsets of which are relatively frequent each month (e.g., an acronym that can refer to many different entities, which may trend at different times). Over time, the word acquires an additional, more stable sense whose frequency does not fluctuate so much from month to month. An example is *BLM*, which has been used to refer to a baseball magazine, a marketing company, a music label, the US Bureau of Land Management, etc., but since 2013 has been consistently associated with the Black Lives Matter movement. This could be considered a 'reductive' meaning change-in-progress, as we start out with multiple competing senses, and one sense gradually comes to dominate without the others having yet died out (see Koch, 2016, Fig. 2).

**Procedure:** Sample eight real words $\{w_1, w_2, ..., w_8\}$ to represent eight different pseudosenses for the pseudoword $\rho$. For each month $m$, draw a multinomial distribution $D^m$ over the first seven sampled words, using a Dirichlet prior with uniform, sparsity-inducing alpha. Replacing each token of a word $w_i, i \in [1, 7]$, insert a token of $\rho$ with success probability $D^m_{w_i}$. Let $w_8$ represent the new, more stable pseudosense, and for each month $m$, insert a token of $\rho$ replacing each token of $w_8$ with success probability $p_{(\rho,w_8,m)}$, which starts low and gradually increases over time.

D1: **Description:** This schema models a word that becomes more frequent over time, but *does not* change its co-occurrence distribution.

**Procedure:** Sample one real word $w$ to represent the meaning of the pseudoword $\rho$. For each token of $w$ that occurs in the synthetic dataset for month $m$, insert a token of $\rho$ replacing the token of $w$ with success probability $p_{(\rho,w,m)}$. $p_{(\rho,w,m)}$ starts relatively low and gradually increases over time.

D2: **Description:** This schema models a word with two senses. One sense is relatively infrequent, but suddenly spikes in frequency (e.g.,

tiple real senses, since the real word we use to represent this pseudosense may itself have multiple senses. There are few words in our dataset with only one sense according to WordNet; so, we restrict our choice to words for which WordNet lists no more than 10 senses. We also require that none of the real words chosen to represent different pseudosenses of a given pseudoword have any senses in common.

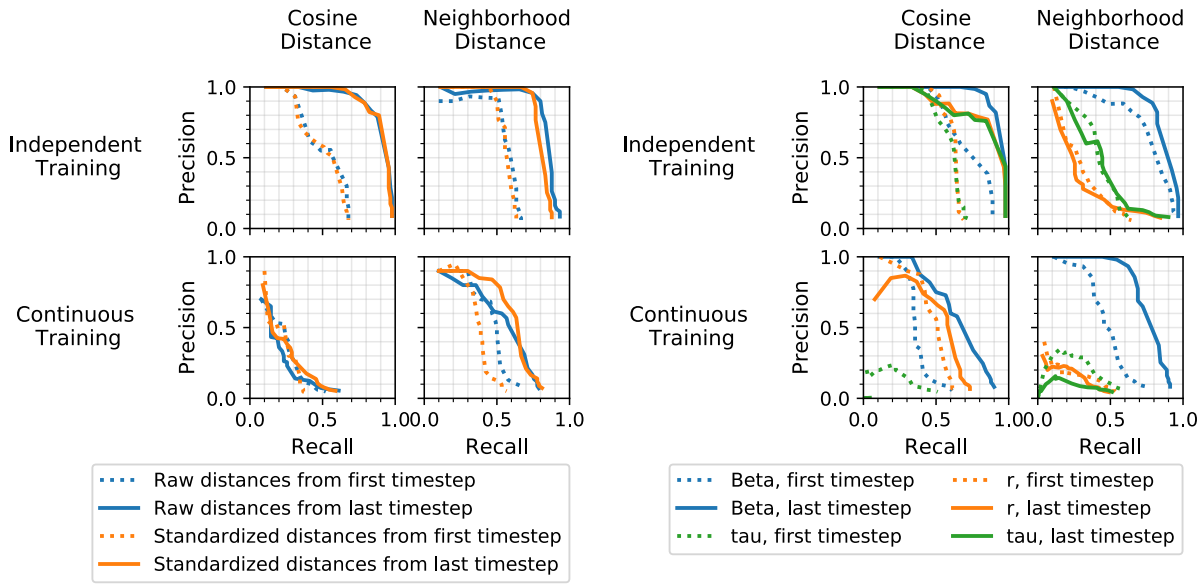


Figure 2: Precision–recall plots for times series approaches for $k$ in range $[0, 1000]$. Left: Change point methods with *Raw* and *Standardized* ($z$) scores. Right: global trend methods including linear regression (*Beta*), Pearson correlation coefficient (*r*), and Kendall rank correlation coefficient (*tau*). Dashed lines use the first timestep as the reference point for comparison while solid lines use the last timestep.

due to a trending topic), before becoming infrequent again.

**Procedure:** Sample two real words $w_1$ and $w_2$ to represent the two pseudosenses. Insert $\rho$ replacing each token of $w_1$ with probability $p_{(\rho,w_1,m)}$, which remains constant throughout the time series, and replacing each token of $w_2$ with probability $p_{(\rho,w_2,m)}$, which starts relatively low, rapidly increases and then rapidly decreases again.

D3: **Description:** This schema models a word with two senses: one is usually relatively infrequent, but spikes in frequency at periodic intervals (i.e., during the same month every year). An example is *turkey*, whose 'poultry' sense tends to be much more frequent around American Thanksgiving and Christmas.

**Procedure:** Sample two real words $w_1$ and $w_2$ to represent the two pseudosenses. Insert $\rho$ replacing each token of $w_1$ with probability $p_{(\rho,w_1,m)}$, which remains constant throughout the time series, and replace each token of $w_2$ with probability $p_{(\rho,w_2,m)}$, which is relatively low for most time-steps but rapidly spikes around the same month each year.

D4: **Description:** Like C3, this schema models words that can refer to many different entities, but in this case, an additional, more stable sense does not emerge.

**Procedure:** Sample seven real words $\{w_1, w_2, ..., w_7\}$ to represent different pseudosenses for $\rho$. For each month $m$, draw a multinomial distribution $D^m$ over these seven words, using a Dirichlet prior with uniform, sparsity-inducing alpha. Replacing each token of a word $w_i, i \in [1, 7]$, insert a token of $\rho$ with success probability $D^m_{w_i}$.

For each of these seven schemas, we create thirty pseudowords (six for each of our five frequency bins), and we vary whether the success probabilities change linearly or logarithmically and the time-steps at which the changes begin and end. In total, we insert 90 pseudowords using Schemas C1–C3, which model genuine semantic changes that we would like to be able to detect, and 120 pseudowords using Schemas D1–D4, which model real changes in words' use statistics but do *not* reflect semantic change. The synthetic dataset also contains 887,926 real words.

### 5.2 Evaluation Results

We evaluate systems by how highly they rank pseudowords from schemas C1–C3 using the Average Precision @ K, which approximates the area under a precision–recall curve over the interval from 0 to K. It is defined as the sum, over every rank $r$ in the top-K list of semantic change

candidates, of the precision at rank $r$ multiplied by the change in recall between ranks $r-1$ and $r$: $AP@K = \sum_{r=1}^{K} P(r)\Delta R(r)$, where $P(r)$ is the percentage of top-$r$ candidates which are pseudowords belonging to Schemas C1–C3, and $R(r)$ is the percentage of all C1–C3 pseudowords that appear in the top-$r$. The results are shown in Table 1 (two-step approach, comparing the first and last time steps) and Table 2 (whole time series). Precision–recall curves for the time series approaches are shown in Figure 2.

**Continuous training does not ensure that embeddings are comparable.** We experiment with three configurations for training the embeddings for the two-step approach: 1) training the embeddings for each time-step independently (ind.); 2) initializing the embeddings for the final time-step with those trained on the first time-step (cont.); and 3) continuous training throughout the whole series, so that the final time-step's embeddings are initialized with the data from all preceding time-steps (cont. whole series).

Continuous training has been used without separate alignment (e.g., Kim et al., 2014) as each time period is a continuation of the embeddings from the previous period. Table 1 shows, however, that alignment is necessary for continuously trained embeddings using the whole series as well as for independent ones when using the cosine distance measure. It is likely that the huge number of training updates in the entire time series causes the embeddings to drift considerably. For the time series approaches, we therefore did not apply the cosine measure without first aligning embeddings.

**Using the whole time series is more effective than comparing the first and the last time steps.** Overall, the approaches using the whole time series (Table 2) are more effective than the two-step approaches; particularly with regard to finding C1 pseudowords and avoiding D4 pseudowords.

**Continuous training provides no benefit for time series approaches.** For the time series approaches, independent training tends to perform better than continuous training (Table 2). The lack of improvement with continuous training is particularly noteworthy as independent training is more computationally efficient than continuous training, since different time periods can be trained in parallel. We did not explore the impact of different hyperparameter choices on continuous training, but note this would introduce another level of complexity in tuning model parameters.

| | ind. | cont. | cont. (whole series) |
|---|---|---|---|
| cosine (unaligned) | 0.00 | 0.32 | 0.00 |
| cosine (aligned) | 0.25 | 0.32 | 0.27 |
| neighbourhood | 0.28 | 0.34 | 0.30 |

Table 1: Average Precision @ 50 on the synthetic dataset of the two-step approach with CBOW

**Different time series approaches are best paired with different similarity measures.** Hamilton et al. (2016a) found that the neighbourhood-based measure tends to assign higher rates of semantic change to nouns, while the cosine measure tends to assign higher rates to verbs. However, they did not compare the overall effectiveness of these methods for semantic change detection.

We find that the neighbourhood-based measure[8] tends to outperform the cosine measure for the change point detection approaches; however, cosine tends to outperform the neighbourhood measure for correlation approaches (see Table 2). For change point detection, standardization of the time series does not have a consistent effect.

**The reference point for comparison matters.** For almost all configurations in Table 2, the AP @ 50 is better when the reference point is the last time-step. Figure 3 shows Recall@K broken down by pseudoword type. For types C1–C3, higher recall is better. Conversely, lower recall is better for types D1–D4, since these model changes that we do *not* consider to be lasting semantic changes. Recall is consistently low for types D1–D4, but strikingly, recall is also low for type C3 when we compare to the first-step. Schema C3 models words whose distributions change drastically from time-step to time-step, but which gradually become more stable as a new, consistently occurring sense emerges. The representation for the first time-step will thus be very different from subsequent representations, such that comparing to the first step is not effective. In contrast, comparing to the first time-step is expected to be more effective in finding words that become less stable over time.

**Correlation-based approaches perform worse than regression or change point detection approaches.** Pearson's correlation coefficient is

[8] We apply this to unaligned embeddings; alignment with orthogonal Procrustes has no effect on this measure.

| Measure | Training | Comparing to the first time-step | | | | | Comparing to the last time-step | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *raw* | $z$ | $\beta$ | $r$ | $\tau$ | *raw* | $z$ | $\beta$ | $r$ | $\tau$ |
| cosine | independent | 0.37 | 0.36 | 0.51 | 0.52 | 0.47 | 0.54 | 0.56 | 0.56 | 0.48 | 0.49 |
| cosine | continuous | 0.17 | 0.18 | 0.34 | 0.40 | 0.02 | 0.13 | 0.15 | 0.45 | 0.34 | 0.00 |
| neighbourhood | independent | 0.47 | 0.50 | 0.48 | 0.23 | 0.35 | 0.53 | 0.56 | 0.56 | 0.20 | 0.30 |
| neighbourhood | continuous | 0.35 | 0.32 | 0.37 | 0.05 | 0.05 | 0.34 | 0.42 | 0.54 | 0.05 | 0.00 |

Table 2: Average Precision @ 50 on the synthetic dataset using time series approaches with CBOW. Change point methods are raw scores (*raw*) and standardized scores ($z$). Global trend methods are linear regression ($\beta$), Pearson correlation coefficient ($r$), and Kendall rank correlation coefficient ($\tau$).

maximized when the *magnitude* of the change between consecutive time periods is consistent over all time periods whereas maximizing Kendall's $\tau$ simply requires the change between consecutive time periods to be of a consistent *sign*. Both correlation measures therefore have particularly poor recall of words that have time periods without a consistent meaning as in the early time periods for pseudowords of type C3.

The $\beta$ value of the linear regression assumes a linear relationship, but is unfortunately sensitive to outliers (Chatterjee and Hadi, 1986), which likely explains why the regression approach has higher recall than change point approaches for schema D4 (Figure 3), in which a stable sense does not emerge. In general, however, the $\beta$ values produced for D4 pseudowords appear to be smaller in magnitude than genuine semantic changes (C1–C3) resulting in average precision measures that generally match or exceed change point approaches (Table 2). Regression is also more straightforward and computationally efficient to calculate than change point measures.

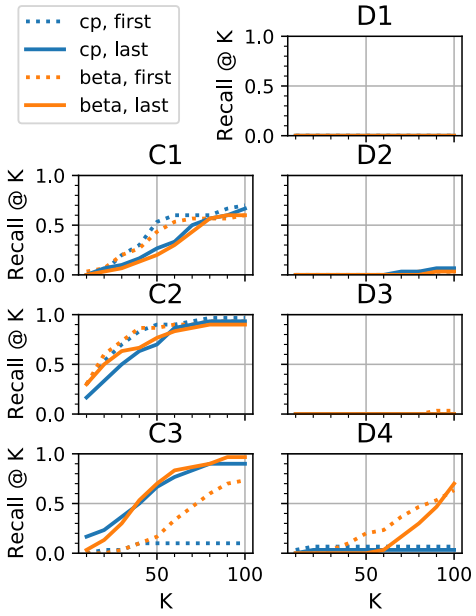


Figure 3: Recall@K of each schema, using independently trained embeddings and the neighbourhood-based measure. 'first'/ 'last' denotes the reference time-step, 'cp' the unstandardized change-point approach, and 'beta' the linear regression approach. C1–C3: higher recall is better, D1–D4: lower is better.

## 6 Results on Empirical Twitter Data

We now apply the approaches on our full empirical Twitter dataset. Table 3 shows the top 10 semantic change candidates using independent, aligned CBOW embeddings. When using continuously trained embeddings, the top-10 lists are similar.

In line with our synthetic results, we find different candidates when comparing to the first time-step or the last, but they appear to represent similar kinds of semantic change. Most have shifted due to associations with named entities. For example, *vine* (Figure 4a) acquired a new sense in January 2013 when the popular short-form video hosting service Vine was launched. Similarly, *ig*, initially shorthand for 'i guess', became shorthand for the social network Instagram as it gained popularity. The embedding for *shawn* shifted significantly around the beginning of 2014, when the singer Shawn Mendes signed a record deal.

There are also words whose embeddings have shifted due to waning associations with prominent named entities, e.g., *vow* was initially associated with The Vow, a high grossing movie released in Feb. 2012, but by the end of the time series it had shifted back towards synonyms like 'pledge' and 'urge'. Likewise the embedding for *temple* initially reflected the popularity of the video game Temple Run but gradually shifted to the word's canonical meaning, and the embedding for *bcs* initially reflected its usage as an acronym for Bowl Championship Series (a selection system in American college football), but then shifted towards

| | |
|---|---|
| **Comparing to first time-step** | vine, temple, unfollowers, favorited, mcm, glo, #ipadgames, shawn, retweeted, vow |
| **Comparing to last time-step** | isis, yasss, bcs, temple, 🐶, mcm, 😐, ig, mila, glo |

Table 3: Top 10 semantic change candidates of the change-point detection approach without standardization, using independently trained and aligned CBOW embeddings and the cosine distance measure.

'bcoz', 'bec', and other forms of 'because' after the selection system was ended in 2013.

There are also examples of neologisms: *mcm* is a lexicalized acronym for 'Man Crush Monday'. This initially referred to the meme of posting about a man one finds attractive each Monday, but then by metonymic extension came to be used to refer to the subject of the post himself. Another example is *glo*, which in the beginning of our data (Figure 4b) occurs mainly in reference to a Nigerian telecommunication company. A shift in its embedding is driven by the sudden emergence of the expression 'glo up', which was coined in August 2013 by rapper Chief Keef in the song "Gotta Glo Up One Day", and later gained traction as an expression to describe an impressive personal transformation.

Finally, there are words whose detected change-points reflect changes in automated activity. For example, the embedding for *yasss* shifts in early 2017 due to a sudden proliferation of tweets automatically posted to users' Twitter accounts by the live video streaming app LiveMe, which all begin with the text 'YASSS It's time for a great show' followed by the title and link to the video stream. Conversely, the detected change for *favorited* (Figure 4c) coincides with a sudden disappearance of automatically generated tweets about favorited YouTube videos.

## 7 Conclusion

In this paper, we presented a new evaluation framework and systematically compared the various choices involved in using word embeddings for semantic change detection. We then applied the approaches to a Twitter dataset spanning 5.5 years. Qualitative analysis found that the top ranked words have undergone genuine semantic change although some of the changes are restricted to social media or to Twitter specifically.

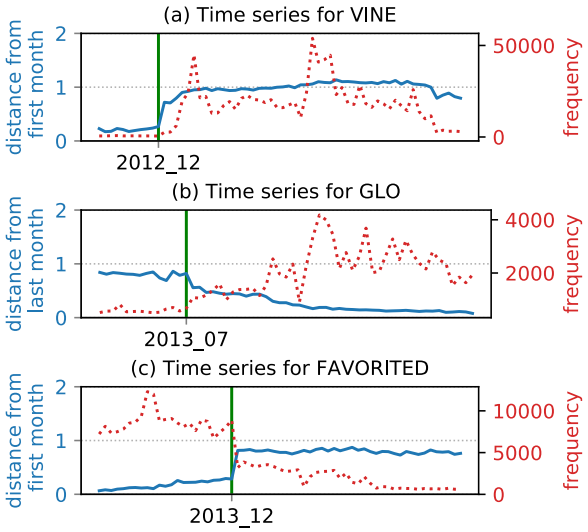


Figure 4: Neighbourhood-based distance (solid blue lines) and frequency (dotted red lines) over time, for three semantic change candidates. Vertical green lines indicate the automatically estimated change-points.

Our framework and dataset can also be used to evaluate approaches not considered in this paper. Moreover, our framework models different semantic change scenarios, and future work could focus on approaches that are able to distinguish between these different scenarios.

## Acknowledgements

This work was supported by The Alan Turing Institute under the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/N510129/1. P.S. was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK EPSRC (grant EP/L016427/1) and the University of Edinburgh. D.N. was supported by Turing award TU/A/000006 and B.McG. by Turing award TU/A/000010 (RG88751). S.A.H. was supported in part by The Volkswagen Foundation.

## References

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 380–389.

Pierpaolo Basile and Barbara McGillivray. 2018. *Discovery Science*, volume 11198 of *Lecture Notes in Computer Science*, chapter Exploiting the Web for Semantic Change Detection. Springer-Verlag.

D. M. Blei and J. D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

Samprit Chatterjee and Ali S. Hadi. 1986. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Mark Graham, Scott A. Hale, and Devin Gaffney. 2014. Where in the world are you? geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, pages 2116–2121.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

Maurice G. Kendall. 1948. *Rank correlation methods*. Griffin, London.

Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1191–1200, New York, NY, USA. ACM.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65.

Peter Koch. 2016. Meaning change and semantic shifts. In Päivi Juvonen and Maria Koptjevskaja-Tamm, editors, *The Lexical Typology of Semantic Shifts*, pages 21–66. De Gruyter Mouton, Berlin/Boston.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2017. Temporal dynamics of semantic relations in word embeddings: an application to predicting armed conflict participants. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1824–1829.

B. McGillivray, S. Hengchen, V. Läteenoja, M. Palma, and A. Vatri. 2019. A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities*.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 259–263.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR) Workshop*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dong Nguyen and Jacob Eisenstein. 2017. A kernel independence test for geographical language variation. *Computational Linguistics*, 43(3):567–592.

Jean-François Quessy, Mériem Saïd, and Anne-Catherine Favre. 2013. Multivariate kendall's tau for change-point detection in copulas. *Canadian Journal of Statistics*, 41(1):65–82.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta.

Martina A. Rodda, Marco S.G. Senaldi, and Alessandro Lenci. 2017. Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. *Italian Journal of Computational Linguistics*, 3:11–24.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, and Frans Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 305–310.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484.

Maja R. Rudolph and David M. Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1003–1011.

Ian Stewart, Dustin Arendt, Eric Bell, and Svitlana Volkova. 2017. Measuring, predicting and visualizing short-term change in word representation and usage in vkontakte social network. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, pages 672–675.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Diachronic Conceptual Change.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Xuri Tang, Weiguang Qu, and Xiaohe Chen. 2016. Semantic change computation: A successive approach. *World Wide Web - Internet & Web Information Systems*, 19(3):375–415.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2018. Short-term meaning shift: an exploratory distributional analysis. *arXiv preprint arXiv:1809.03169*.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pages 35–40.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 673–681.